Learning and inference in a nonequilibrium Ising model with hidden nodes

Benjamin Dunn*
The Kavli Institue for Systems Neuroscience, NTNU, 7030 Trondheim

Yasser Roudi†
The Kavli Institue for Systems Neuroscience, NTNU, 7030 Trondheim and
NORDITA, KTH Royal Institute of Technology and Stockholm University, 10691 Stockholm, Sweden

We study inference and reconstruction of couplings in a partially observed kinetic Ising model. With hidden spins, calculating the likelihood of a sequence of observed spin configurations requires performing a trace over the configurations of the hidden ones. This, as we show, can be represented as a path integral. Using this representation, we demonstrate that systematic approximate inference and learning rules can be derived using dynamical mean-field theory. Although naive mean-field theory leads to an unstable learning rule, taking into account Gaussian corrections allows learning the couplings involving hidden nodes. It also improves learning of the couplings between the observed nodes compared to when hidden nodes are ignored.

PACS numbers: 02.50.Tt,05.10.-a,89.75.Hc,75.10.Nr

I. INTRODUCTION

Within the statistical mechanics community, a significant body of recent work deals with parameter learning in statistical models. This work typically focuses on the canonical setting of the inverse Ising problem: given a set of configurations from an equilibrium [1], or nonequilibrium [2] Ising model, how can we find the interactions between the spins? These inverse problems can be difficult to solve using brute force approaches [3] and thus for practical purposes, the results of such theoretical works offering approximate learning and inference methods are of great importance. The recent interest in using statistical models to study high-throughput data from neural [4–6], genetic [7] and financial [8] networks further emphasizes the necessity of developing efficient approximate inference methods.

Despite its practical relevance, however, little has been done on learning and inference in models with hidden spins using statistical physics. Learning the parameters of a statistical model in the presence of hidden nodes and inferring the state of hidden variables lie at the heart of difficult problems in statistics and machine learning. It is well known that for statistical modeling of data, hidden variables most often improve model quality [9, 10]. Furthermore, when used for network reconstruction and graph identification, ignoring hidden nodes may cause the model to incorrectly identify/remove couplings between observed nodes when such couplings do not exist/actually exist. For these reasons, it is quite important to study models with hidden nodes, both from a theoretical perspective and for applications to real data where usually only a fraction of the system is observeable [4, 5, 7].

In this paper, we study the reconstruction of the couplings in a nonequilibrium Ising model with hidden spins and the inference of the state of these spins. Theoretically, the couplings can be reconstructed by calculating and maximizing the likelihood of the data which involves performing a trace over the configurations of the hidden spins at all times: an impossible task even for moderately sized data sets and networks. This is, in fact, a general problem, not only for a nonequilibrium Ising model, but for any kinetic model with hidden variables. It is, however, natural to use a path integral approach for doing this trace, and this is what we do here. We show that efficient approximate learning and inference can be developed by evaluating the log-likelihood in this way and employing mean-field type approximations.

In what follows, after formulating the problem, we first evaluate the likelihood of the data using the saddle point approximation to the path integral. We find that the result allows inference of the state of hidden variables if the couplings are known but cannot be used for learning the couplings. We then perform a Gaussian correction to this saddle point approximation to derive Thouless-Anderson-Palmer (TAP) like equations [11] for our problem. Our numerical results show that successul learning of the couplings, even those that involve hidden nodes, can be achieved using the corrected learning rules. We find that with enough data and using these equations even hidden-to-hidden connections can be reconstructed.

Although in this paper our focus is on the paradigmatic case of the Ising model, the approach taken here is quite general and can be adapted to many other popular statistical models.

II. FORMULATION OF THE PROBLEM

Consider a network of binary spin variables subject to synchronous update dynamics. Suppose some of these nodes are observed for T time steps, while the other nodes are hidden. We denote the state of the observed spins at

* benjamin.dunn@ntnu.no
† yasser.roudi@ntnu.no

time t by $\boldsymbol{s}(t) = \{s_i(t)\}$ and the hidden ones by $\boldsymbol{\sigma} = \{\sigma_a(t)\}$. For clarity, we always use the subscripts $i, j, \ldots$ for the observed and $a, b, \ldots$ for the hidden ones. We consider the following transition probability

$$p[\{\boldsymbol{s},\boldsymbol{\sigma}\}(t+1)|\{\boldsymbol{s},\boldsymbol{\sigma}\}(t)] = \exp[\textstyle\sum_i s_i(t+1)g_i(t) + \sum_a \sigma_a(t+1)g_a(t)]Z(t)^{-1} \tag{1a}$$

$$Z(t) = \prod_{i,a} 2\cosh[g_i(t)]\ 2\cosh[g_a(t)] \tag{1b}$$

where $g_i(t)$ and $g_a(t)$ are the fields acting on observed spin i and hidden spin a at time t, respectively

$$g_i(t) = \textstyle\sum_j J_{ij}s_j(t) + \sum_b J_{ib}\sigma_b(t) \tag{2a}$$

$$g_a(t) = \textstyle\sum_j J_{aj}s_j(t) + \sum_b J_{ab}\sigma_b(t), \tag{2b}$$

J_{ij}, J_{ai}, J_{ia} and J_{ab} are the observed-to-observed, observed-to-hidden, hidden-to-observed and hidden-to-hidden connections. These couplings need not be symmetric and, thus, the network may never reach equilibrium. Although, here we consider the case of zero external field, our derivations follow also for nonzero fields.

Given this model, we would like to answer the following questions: What can we say about the state of hidden spins from the observations? How can we learn the various couplings in the network?

Answering these questions requires finding the likelihood of the data. Optimizing the likelihood with respect to Js yields the maximum likelihood value of the couplings. The posterior over the state of hidden nodes given the observed nodes is

$$p[\{\boldsymbol{\sigma}(t)\}_{t=1}^T|\{\boldsymbol{s}(t)\}_{t=1}^T] = \frac{p[\{\boldsymbol{\sigma}(t)\}_{t=1}^T, \{\boldsymbol{s}(t)\}_{t=1}^T]}{p[\{\boldsymbol{s}(t)\}_{t=1}^T]}$$

which also requires calculating the likelihood.

The likelihood of the observed spin configurations under the model Eq. 1 is

$$p[\{\boldsymbol{s}(t)\}_{t=1}^T] = \mathrm{Tr}_\sigma \prod_t p[\{\boldsymbol{s},\boldsymbol{\sigma}\}(t+1)|\{\boldsymbol{s},\boldsymbol{\sigma}\}(t)] \tag{3}$$

and involves a trace over $\{\boldsymbol{\sigma}(1), \ldots, \boldsymbol{\sigma}(T)\}$ which is difficult to do.

To perform this trace analytically, we consider the following functional

$$\mathcal{L}[\psi] \equiv \log \mathrm{Tr}_\sigma \prod_t e^{\sum_a \psi_a(t)\sigma_a(t)} p[\{\boldsymbol{s},\boldsymbol{\sigma}\}(t+1)|\{\boldsymbol{s},\boldsymbol{\sigma}\}(t)]. \tag{4}$$

which, using the definition of the transition probability from Eq. 1, can be written as

$$\mathcal{L}[\psi] = \log \mathrm{Tr}_\sigma \exp\left[Q[\boldsymbol{s},\boldsymbol{\sigma}] + \sum_{a,t}\psi_a(t)\sigma_a(t)\right] \tag{5}$$

$$Q = \sum_{i,t} s_i(t+1)g_i(t) + \sum_{a,t}\sigma_a(t+1)g_a(t) - \sum_{i,t}\log 2\cosh[g_i(t)] - \sum_{a,t}\log 2\cosh[g_a(t)].$$

The log-likelihood of the data is recovered at $\psi \to 0$. Furthermore, $\mathcal{L}[\psi]$ acts as the generating functional for the posterior over the hidden spins: $m_a(t)$, the mean magnetization of the hidden spin a at time t given the data, can be found using

$$m_a(t) = \lim_{\psi\to 0}\mu_a(\psi, t), \qquad \mu_a(\psi,t) = \frac{\partial\mathcal{L}[\psi]}{\partial\psi_a(t)}. \tag{6}$$

Higher-order derivatives of $\mathcal{L}$ with respect to ψ yield higher-order correlation functions of the hidden spins.

To develop mean-field approximations, we first perform the trace in Eq. 5 using the standard approach for dealing with a nonlinear action in the local fields [12, 13]: we write the term inside the log in Eq. 5 as a path integral over $g_i(t)$ and $g_a(t)$ enforcing their definitions in Eq. 2 by inserting $\delta[g_i(t) - \sum_j J_{ij}s_j(t) - \sum_b J_{ib}\sigma_b(t)]$ and $\delta[g_a(t) - \sum_j J_{aj}s_j(t) - \sum_b J_{ab}\sigma_b(t)]$ in the integral. Using the integral representation of the delta functions, $\mathcal{L}$ can be written as

$$\mathcal{L}[\psi] = \log \int D\mathcal{G}\exp[\Phi] \tag{7a}$$

$$\Phi = \log \mathrm{Tr}_{\boldsymbol{\sigma}} \exp[Q + \Delta + \sum_{a,t}\psi_a(t)\sigma_a(t)] \tag{7b}$$

where $\mathcal{G} = \{g_i, \hat{g}_i, g_a, \hat{g}_a\}$. Δ, as well as $\hat{g}_i$ and $\hat{g}_a$, come from the integral representation of the $\delta(\cdot)$ functions used to enforce the definitions in Eq. 2

$$\Delta = \sum_{i,t} i\hat{g}_i(t)[g_i(t) - \sum_j J_{ij}s_j(t) - \sum_b J_{ib}\sigma_b(t)] + \sum_{a,t} i\hat{g}_a(t)[g_a(t) - \sum_j J_{aj}s_j(t) - \sum_b J_{ab}\sigma_b(t)] \tag{8}$$

The crucial thing about this rewriting of $\mathcal{L}$ is that the resulting action Φ is now linear in σ and we can thus easily perform the trace over σ. The ability to perform the trace here comes at the cost of the high-dimensional integral over $\mathcal{G}$ in Eq. 7a. The integral, however, can be approximated using the saddle point approximation and the corrections to the saddle point.

III. SADDLE-POINT APPROXIMATION

The saddle point values of $g_a, g_i, \hat{g}_a$ and $\hat{g}_i$ can be derived by putting the derivatives of Φ with respect to these variables to zero. For $\psi = 0$, the saddle point equations read

$$i\hat{g}_i(t) = \tanh[g_i(t)] - s_i(t+1) \tag{9a}$$

$$i\hat{g}_a(t) = \tanh[g_a(t)] - m_a(t+1) \tag{9b}$$

$$g_i(t) = \textstyle\sum_j J_{ij}s_j(t) + \sum_b J_{ib}m_b(t) \tag{9c}$$

$$g_a(t) = \textstyle\sum_j J_{aj}s_j(t) + \sum_b J_{ab}m_b(t) \tag{9d}$$

in which $m_a(t)$ satisfy the self-consistent equations

$$m_a(t) = \tanh\left[g_a(t-1) - i\textstyle\sum_j \hat{g}_j(t)J_{ja} - i\sum_b \hat{g}_b(t)J_{ba}\right] \tag{10}$$

Let us take a moment to describe the physical meaning of these equations. The saddle point equations for g_i and g_a, Eq. 9c and 9d, are just the mean values of the local fields at the mean magnetizations $m_a(t)$ of the hidden variables (see Eq. 2). Examining Eq. 9a and 9b, one can identify $-i\hat{g}_i(t)$ and $-i\hat{g}_a(t)$ as errors back propagated in time. The self-consistent equations for m_a, Eq. 10, take the form of the usual naive mean-field equations except that back propagated errors are taken into account.

Given a sequence of configurations from the observed spins, solving Eq. 10 yields the mean magnetization of the hidden ones given this data. The derivatives of Φ with respect to the Js, evaluated at the saddle point solutions, give the maximum likelihood learning rules for the couplings within this saddle point approximation.

IV. GAUSSIAN CORRECTIONS

The saddle-point equations and learning rules can be improved by Gaussian corrections, that is by evaluating the contributions to the path integral from the Gaussian fluctuations around the saddle [14, 15]. This will lead to what can be called as TAP equations for our problem. In what follows, by doing this on the Legendre transform of $\mathcal{L}$,

$$\Gamma[\mu] = \mathcal{L} - \sum_{a,t} \psi_a(t)\mu_a(t) \tag{11}$$

and using the equation of state,

$$-\psi_a(t) = \frac{\partial \Gamma[\mu]}{\partial \mu_a(t)} \tag{12}$$

we correct the naive mean-field equations for our problem with the hidden spins.

Denoting the saddle point value of $\mathcal{L}$ by $\mathcal{L}_0$, and performing a Legendre transform we have (see Appendix A)

$$\Gamma_0[\mu] = \sum_{i,t}[s_i(t+1)g_i(t) - \log 2\cosh(g_i(t))] \tag{13}$$
$$+ \sum_{a,t}[\mu_a(t+1)g_a(t) - \log 2\cosh(g_a(t))] + \sum_{a,t} S[\mu_a(t)]$$

where $S[\mu_a(t)]$ is the entropy of a free spin with magnetization $\mu_a(t)$, and g_i and g_a are as in Eq. 2 with $\sigma_b(t)$ replaced by $\mu_b(t)$. Using Eq. 12 and putting $\psi = 0$, we recover Eq. 10, as we should.

The Gaussian corrections can be performed as follows; for details see Appendix B. We first perform the Gaussian integral around the saddle point of $\mathcal{L}$. This modifies $\mathcal{L}_0$ to

$$\mathcal{L}_1 = \mathcal{L}_0 - \frac{1}{2}\log\det[\partial^2 \mathcal{L}] \tag{14}$$

where $\partial^2\mathcal{L}$ is the Hessian, containing the second derivatives of $\mathcal{L}$ with respect to g and $\hat{g}$, calculated at their saddle point values. Keeping terms quadratic in J we then derive the modified equations for $m_a(t)$ by first calculating $\partial\mathcal{L}_1/\partial\psi_a(t)$. Expressing $\mathcal{L}_1$ in terms of the new $m_a(t)$, again keeping terms quadratic in J, and finally performing the Legendre transform, we get

$$\Gamma_1[m] = \Gamma_0[m] - \frac{1}{2}\sum_{i,a,t}[1-\tanh^2 g_i(t)]J_{ib}^2[1-m_b^2(t)]$$
$$-\frac{1}{2}\sum_{a,b,t}[m_a^2(t+1) - \tanh^2 g_a(t)]J_{ab}^2[1-m_b^2(t)]. \tag{15}$$

where now, g_i and g_a are as in Eq. 2 but with m_a instead of σ_a. Eq. 15 together with the equation of state at $\psi \to 0$,

$$\frac{\partial\Gamma_1[m]}{\partial m_a(t)} = 0, \tag{16}$$

yields the new self-consistent equation for $m_a(t)$. The couplings can then be found by optimizing Γ_1 with respect to the couplings.

In sum, the corrected TAP-learning will be the following EM like iterative algorithm: at each step, solve the equations for m_a derived from Eq. 16 given the current value of the couplings, and then adjust the couplings proportional to $\partial\Gamma_1/\partial J$.

V. NUMERICAL RESULTS

We evaluated the saddle point and TAP equations on data generated from a network of $N = 100$ spins with the dynamics in Eq. 1 and with the couplings drawn independently from a Gaussian distribution with variance J_1^2/N. We first studied how much the saddle point Eq. 10 and TAP equations Eq. 16 tell us about the state of hidden spins if we know all the couplings. Fig. 1 A and B show the distributions of $m_a(t)$ when 20% of the network is hidden. The solutions to the TAP equations are more pushed towards $+1$ and -1 compared to the saddle point. Using a simple estimator $\hat{\sigma}_a(t) = \mathrm{sgn}(m_a(t))$ to infer the state of hidden spin a and time t, Fig. 1C shows that the two approximations perform equally well in inferring the state of hidden spins. The effect of TAP corrections becomes very important in reconstructing the couplings. We found that the saddle point learning was always unstable leading to divergent couplings. However, including the TAP corrections dramatically changed this. Fig. 2 shows an example of this when 10% of the spins were hidden. In this case, good reconstruction of all couplings could be achieved.

Although for a proportionally small number of hidden nodes we observed good reconstruction, when the hidden part was large, for a constant data length, we found that the algorithm starts to become unstable: the root mean squared error for the inferred hidden-to-hidden connections increases after a few iterations and the others follow. This instability, however, seems to arise from our attempt to learn the hidden-to-hidden connections.

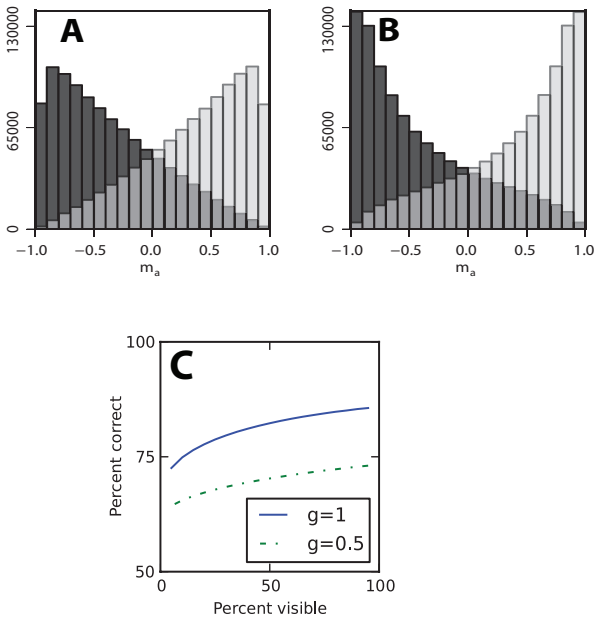

FIG. 1. (Color online) Solutions for m_a. (A) and (B) show histograms of $m_a(t)$ found from solving saddle point and TAP equations, Eqs. 10, 16, given the correct couplings. Here we had $J_1 = 0.7$, 20% hidden spins and a data length of $T = 10^5$. The dark gray bars show the histogram of the values for which $\sigma_a(t)$ was actually $+1$ while the transparent ones show when it was actually -1. (C) shows the percent correct if we use $\hat{\sigma}_a(t) = \mathrm{sgn}(m_a(t))$ to infer the state of $\sigma_a(t)$ vs. the percentage of the visible part. TAP and saddle point results were virtually indistinguishable.

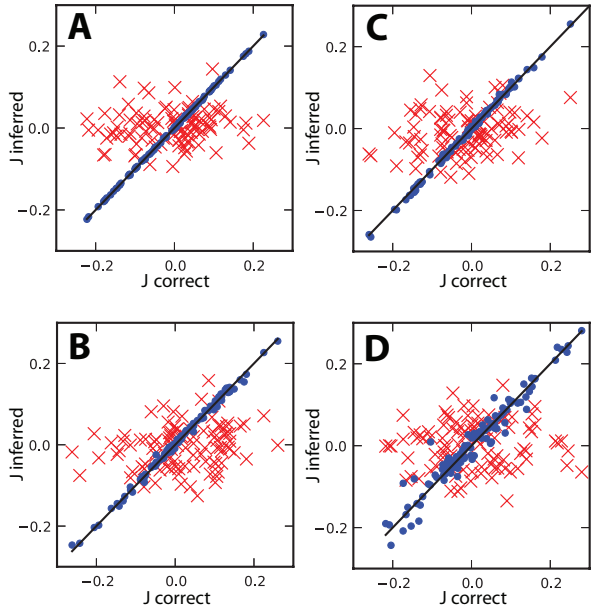

FIG. 2. (Color online) Reconstructing the full network. Scatter plots showing the inferred couplings (blue dots) versus the correct ones in a network of $N = 100$ total spins, with 10 hidden, $J_1 = 1$ and data length of $T = 10^6$. Red crosses show the initial values of the couplings used for learning. A, B, C, and D show observed-to-observed, observed-to-hidden, hidden-to-observed and hidden-to-hidden couplings, respectively.

For an architecture where there is no hidden-to-hidden couplings, that is a dynamical semi-restricted Boltzmann machine [16], or when hidden-to-hidden couplings exist in the original network but are ignored during learning (i.e. put to zero), the algorithm allows recovering the other connections even for very large hidden sizes. This is shown in Fig. 3.

In the examples above, we have used the right number of hidden nodes when learning the couplings. In practice, however, the number of hidden nodes may be unknown. We thus wondered if the number of hidden nodes can be estimated from the data. In Fig. 4, we plot the objective function Γ_1/N for the data used in Fig. 2, where N is the total number of visible nodes, that is, the sum of the observed nodes, and an unknown number M of hidden nodes. As can be seen in Fig. 4, varying M, at the correct number of hidden units a peak in Γ_1/N is seen and also the reconstruction error of the visible-to-visible connectivity reaches a minimum. This was the case, even when we constrained the hidden-to-hidden connections to zero during learning. However, the peak in the normalized cost function was sharper, and the minmum achieved for the reconstruction lower, when hidden-to-hidden connections were also learned.

VI. DISCUSSIONS

In this paper we described an approach for learning and inference in a nonequilibrium Ising model when it is partially observed, that is, when the history of spin configurations are known only for some spins and not for the others. Treating the log-likelihood of the data as a path integral over the fields acting on hidden and observed spins in a non-equilibrium Ising model, we approximated it first using saddle point approximations, and then by considering TAP corrections to it.

In usual statistical mechanics settings, TAP equations can be derived in several ways: using Plefka expansion [17], information-geometric arguments [18] or by calculating the Gaussian fluctuations around the saddle point [14, 15]. Here we took an approach similar to the latter case except that we did not use the Hubbard-Stratonovich transformation. It would be quite interesting to see what the other approaches yield as TAP corrections. Tyrcha and Hertz [19] have also studied the problem of hidden spins, using a different approach. Their equations agree with ours at the saddle point level but differ at the TAP level.

In terms of performance, saddle point equations gave results that were similar to TAP for inferring the state of hidden spins, but were very unstable for learning the couplings. Using TAP equations, learning all the couplings, even the hidden-to-hidden ones, was possible

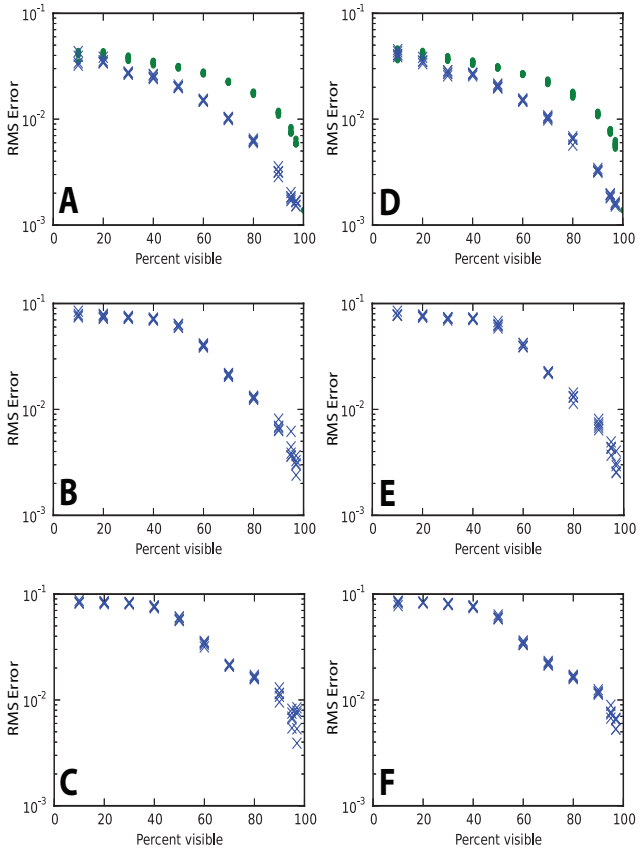

FIG. 3. (Color online) Root mean squared error of the couplings. (A)-(C) show the rms error for the observed-to-observed, hidden-to-observed and observed-to-hidden couplings using TAP learning when the original model did not have hidden-to-hidden connections; the other couplings were drawn independently from a Gaussian. (D)-(F) show the same but when the hidden-to-hidden connections were present in the original network but were ignored (i.e. put to zero) when learning the other couplings. In (A) and (D) the green dots show the inferred couplings by maximizing the likelihood with hidden nodes completely ignored. Here we used $J_1 = 1, N = 100, T = 10^6$, and five different random realizations of the connectivity.

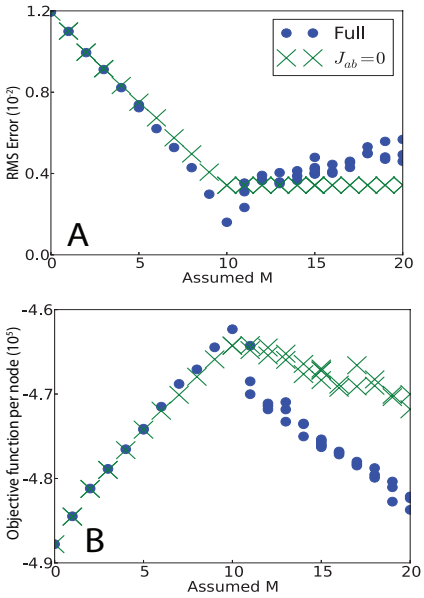

FIG. 4. (Color online) Estimation of the number of hidden nodes. (A) the RMS error of the observed-to-observed couplings inferred using TAP learning for different values of assumed hidden units, M. The couplings are inferred either assuming no hidden-to-hidden connectivity (green x) or by trying to learn all couplings (blue circles). (B) the value of the objective function per inferred node for different numbers of assumed hidden units. The data used here is the same as that of Fig. 2, that is in reality there were 10 hidden nodes which is where the normalized objective function shows maximum and the RMS error shows a minimum.

when the hidden part was proportionally small. By ignoring hidden-to-hidden couplings, even when they were present in the network, we could achieve good performance in recovering all the other couplings, even with half the network hidden. This also improved recovering the observed-to-observed couplings compared to when hidden spins were ignored all together. Our numerical results also indicate that even the number of hidden units can be estimated by optimization of the objective function. This is of particular practical revelance as often the true size of the network is unknown. In Fig. 4, both the objective function and the reconstruction quality increase as we add hidden nodes suggesting that even a wrong number of hidden nodes can improve the reconstruction of observed-to-observed couplings. This can be explained by the fact that the added hidden variables, even if wrong in number, to some degree can explain the correlations induced by the hidden spins in the real network.

An important future direction, in our view, is to study how averaging the likelihood over some couplings, in particular the hard to infer hidden-to-hidden ones, given a prior over them, will influence the reconstruction of the other couplings. There is a large body of work where the path integral approach is used to study the disorder-averaged behavior of disordered systems [12, 20]. The approach that we developed here can be easily combined with this work to average out some of the couplings.

VII. ACKNOWLEDGEMENT

We are most grateful to Joanna Tyrcha and John Hertz for constructive comments and for discussing their work with us, and Claudia Battistin for comments on the paper. Benjamin Dunn acknowledges the hospitality of NORDITA. We also acknowledge the computing resources provided by the University of Oslo and the Norwegian metacenter for High Performance Computing (NOTUR), Abel cluster.

Appendix A: Legendre transform

Here we derive Eq. 13 in the main text. Starting from Eq. 7, we have

$$\mathcal{L}[\psi] = \log \int D\mathcal{G} \exp[\Phi], \tag{I-1a}$$

$$\Phi = \sum_{i,t} s_i(t+1) g_i(t) + i \sum_{i,t} \hat{g}_i(t) \Big[g_i(t) - \sum_j J_{ij} s_j(t) \Big] + i \sum_{a,t} \hat{g}_a(t) \Big[g_a(t) - \sum_i J_{ai} s_i(t) \Big] - \sum_{i,t} \log 2\cosh[g_i(t)]$$
$$- \sum_{a,t} \log 2\cosh[g_a(t)] + \sum_{a,t} \log 2\cosh \Big[g_a(t-1) - i \sum_b J_{ba} \hat{g}_b(t) - i \sum_i J_{ia} \hat{g}_i(t) + \psi_a(t) \Big] \tag{I-1b}$$

$$\mu_a(t) = \frac{\partial \Phi}{\partial \psi_a(t)} = \tanh \big[g_a(t-1) - i \sum_j \hat{g}_j(t) J_{ja} - i \sum_b \hat{g}_b(t) J_{ba} + \psi_a(t) \big], \tag{I-1c}$$

with the saddle point being at Eq. 9, with μ_a instead of m_a. Applying the following identity to the last sum in Eq. I-1b

$$\log 2\cosh[x] = S[\tanh[x]] + x \tanh[x] \tag{I-2}$$

$$S[x] \equiv -\frac{1+x}{2} \log \left[\frac{1+x}{2} \right] - \frac{1-x}{2} \log \left[\frac{1-x}{2} \right] \tag{I-3}$$

and using Eq. I-1c and the saddle point values for g_i and g_a, we obtain

$$\Phi_0 = \sum_{i,t} [s_i(t+1) g_i(t) - \log 2\cosh[g_i(t)]]$$
$$+ \sum_{a,t} [\mu_a(t+1) g_a(t) - \log 2\cosh[g_a(t)]]$$
$$+ \sum_{a,t} \mu_a(t) \psi_a(t) + \sum_{a,t} S[\mu_a(t)] \tag{I-4}$$

At the saddle point we have $\mathcal{L}_0 = \Phi_0$ and using $\Gamma_0 = \mathcal{L}_0 - \sum_{a,t} \psi_a(t) \mu_a(t)$ we arrive at Eq. 13.

Appendix B: Corrections to the saddle-point likelihood

To include the Gaussian corrections in Eq. 15 we first calculate the Gaussian fluctuations around the saddle point value of $\mathcal{L}_0$. Given Eq. I-1b, we have

$$A_{ij}^{tt'} \equiv \frac{\partial^2 \Phi}{\partial g_i(t) \partial g_j(t')} = -\delta_{ij} \delta_{tt'} [1 - \tanh^2[g_i(t)]]$$

$$B_{ij}^{tt'} \equiv \frac{\partial^2 \Phi}{\partial \hat{g}_i(t) \partial \hat{g}_j(t')} = -\sum\nolimits_a J_{ia} J_{ja} [1 - \mu_a^2(t)] \delta_{tt'}$$

$$C_{ab}^{tt'} \equiv \frac{\partial^2 \Phi}{\partial g_a(t) \partial g_b(t')} = -\delta_{ab} \delta_{tt'} \Big[\mu_a^2(t+1) - \tanh^2[g_a(t)]] \Big]$$

$$D_{ab}^{tt'} \equiv \frac{\partial^2 \Phi}{\partial \hat{g}_a(t) \partial \hat{g}_b(t')} = -\sum\nolimits_c J_{ac} J_{bc} [1 - \mu_c^2(t)] \delta_{tt'}$$

$$E_{ib}^{tt'} \equiv \frac{\partial^2 \Phi}{\partial \hat{g}_i(t) \partial \hat{g}_b(t')} = -\sum\nolimits_a J_{ia} J_{ba} [1 - \mu_a^2(t)] \delta_{t,t'}$$

$$F_{ib}^{tt'} = \frac{\partial^2 \Phi}{\partial \hat{g}_i(t) \partial g_b(t')} = -i J_{ib} [1 - \mu_b^2(t)] \delta_{t-1,t'}$$

$$\frac{\partial^2 \Phi}{\partial g_a(t) \partial \hat{g}_b(t')} = i \delta_{ab} \delta_{tt'} \underbrace{-i J_{ba} [1 - \mu_a^2(t+1)] \delta_{t',t+1}}_{G_{ab}}$$

and $\partial^2 \Phi / \partial g_i(t) \partial \hat{g}_j(t') = i \delta_{ij} \delta_{tt'}$leading to

$$\left[\begin{array}{cc:cc|cc:cc}
A^{tt} & i\mathbb{1} & 0 & 0 & 0 & 0 & 0 & 0 \\
i\mathbb{1} & B^{tt} & 0 & E^{tt} & 0 & 0 & 0 & 0 \\ \hdashline
0 & 0 & C^{tt} & i\mathbb{1} & 0 & [F^{t't}]^T & 0 & G^{tt'} \\
0 & [E^{tt}]^T & i\mathbb{1} & D^{tt} & 0 & 0 & 0 & 0 \\ \hline
0 & 0 & 0 & 0 & A^{t't'} & i\mathbb{1} & 0 & 0 \\
0 & 0 & F^{t't} & 0 & i\mathbb{1} & B^{t't'} & 0 & E^{t't'} \\ \hdashline
0 & 0 & 0 & 0 & 0 & 0 & C^{t't'} & i\mathbb{1} \\
0 & 0 & [G^{tt'}]^T & 0 & 0 & [E^{t't'}]^T & i\mathbb{1} & D^{t't'}
\end{array}\right] \tag{II-2}$$

as part of the Hessian for t and $t' = t+1$, and the same structure gets repeated for other times. The first to the fourth rows/columns correspond to derivatives with respect to $g_i, \hat{g}_i, g_a, \hat{g}_a$ all at t and the fourth to eighth

rows/columns at $t' = t + 1$.

Denoting the matrix of the blocks on the diagonal part of the Hessian as α and the rest as β, we write

$$\begin{aligned}&\log\det(\alpha+\beta) = \log\det(\alpha) + \log\det[I+\alpha^{-1}\beta]\\ &= \log\det(\alpha) + \mathrm{Tr}\log[I+\alpha^{-1}\beta] \qquad \text{(II-3)}\\ &\approx \log\det(\alpha) + \mathrm{Tr}[\alpha^{-1}\beta] + \frac{1}{2}\mathrm{Tr}\{[\alpha^{-1}\beta]^2\}\} + \cdots\end{aligned}$$

Assuming that the couplings are random and independent with mean of order $1/N$ and a standard deviation of $J_1/\sqrt{N}$ as typically assumed in mean-field models of spin glasses [21], $\log\det(\alpha)$ will be of quadratic order in J_1. Since α is block diagonal, so will be α^{-1} and therefore $\mathrm{Tr}[\alpha^{-1}\beta] = 0$. Ignoring the second trace as contributing higher order terms, we thus approximate the determinant of the Hessian matrix by considering only the block diagonal terms

$$\det(\alpha) = \prod_t \det\{A(t)B(t)+\mathbb{1}\}\det\{C(t)D(t)+\mathbb{1}\} \quad \text{(II-4)}$$

leading to the following estimate of the contribution of the Gaussian fluctuations

$$\begin{aligned}\delta\mathcal{L} &\approx -\frac{1}{2}\sum_t \log\det\{A(t)B(t)+\mathbb{1}\} \qquad \text{(II-5)}\\ &-\frac{1}{2}\log\det\{C(t)D(t)+\mathbb{1}\}\\ &\approx -\frac{1}{2}\sum_{t,i}\left\{[1-\tanh^2[g_i(t)]]\sum_b J_{ib}^2[1-\mu_b^2(t)]\right\}\\ &-\frac{1}{2}\sum_{t,a}\left\{[\mu_a^2(t+1)-\tanh^2(g_a(t))]\sum_b J_{ab}^2[1-\mu_b^2(t)]\right\}\end{aligned}$$

With these Gaussian fluctuations taken into account, the corrected value for the mean magnetization is

$$m_a(t) = \frac{\partial\mathcal{L}}{\partial\psi_a(t)} = \mu_i(t) + l_a(t) \qquad \text{(II-6)}$$

$$l_a(t) = \frac{\partial\delta\mathcal{L}}{\partial\psi_a(t)} \qquad \text{(II-7)}$$

We now express Γ_0 to the quadratic order in J in terms of m_a instead of μ_a. With some algebra at $\psi \to 0$ we have

$$\begin{aligned}&\Gamma_0[m] = \sum_{i,t}[s_i(t+1)g_i(t) - \log 2\cosh(g_i(t))]\\ &+\sum_{a,t}[m_a(t+1)g_a(t) - \log 2\cosh(g_a(t))] + \sum_{a,t} S[m_a(t)]\\ &-\sum_{i,t,a}[s_i(t+1) - \tanh(g_i(t))]J_{ia}l_a(t)\\ &+\sum_{a,t} l_a(t+1)[\tanh^{-1}[m_a(t+1)] - g_a(t)]\\ &-\sum_{a,t,b}[m_a(t+1) - \tanh[g_a(t)]J_{ab}l_b(t)\end{aligned}$$

Where we have abused the notation here, using g_a and g_i to indicate the fields calculated at $m_a(t)$ instead of $\mu_a(t)$. Noting that l_a itself is quadratic in J, the terms that involve l_a above are all higher order, and can therefore be ignored. This, combined with the expression for $\delta\mathcal{L}$ in Eq. II-5 yield Eq. 15 in the main text.

[1] H. J. Kappen and F. B. Rodriguez, Neur. Comp. **10**, 1137 (1998); T. Tanaka, Phys. Rev. E **58**, 2302 (1998); Y. Roudi, E. Aurell, and J. Hertz, Front. Comp. Neur. **3**, 22 (2009); S. Cocco and R. Monasson, Phys. Rev. Lett. **106**, 090601 (2011); F. Ricci-Tersenghi, J. Stat. Mech.: Theory and Exp. , P08015 (2012); H. C. Nguyen and J. Berg, J. Stat. Mech.: Theory and Exp. , P03004 (2012); Phys. Rev. Lett. **109**, 050602 (2012).

[2] Y. Roudi and J. Hertz, Phys. Rev. Lett. **106**, 048702 (2011); M. Mezard and J. Sakellariou, J. Stat. Mech.: Theory and Exp. (2011); H.-L. Zeng, E. Aurell, M. Alava, and H. Mahmoudi, Phys. Rev. E. **83**, 041135 (2011); P. Zhang, J. Stat. Phys. , 1 (2012).

[3] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, Cognitive Science **9**, 147 (1985).

[4] E. Schneidman, M. Berry, R. Segev, and W. Bialek, Nature **440**, 1007 (2006).

[5] J. Shlens, G. Field, J. Gauthier, M. Grivich, D. Petrusca, A. Sher, A. Litke, and E. Chichilnisky, J. Neurosci. **26**, 8254 (2006).

[6] S. Cocco, S. Leibler, and R. Monasson, Proc. Natl. Acad. Sci. USA **106**, 14058 (2009).

[7] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. Fedoroff, Proc. Natl. Acad. Sci. USA **103** (2006).

[8] T. Bury, Physica A **392**, 13751385 (2012).

[9] C. Bishop, in *Learning in graphical models*, edited by M. I. Jordan (MIT Press, 1999).

[10] J. Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge Univ. Press, 2000).

[11] D. J. Thouless, P. W. Anderson, and R. G. Palmer, Philosophical Magazine **35**, 593 (1977).

[12] A. Coolen, in *Neuro-Informatics and Neural Modelling*, Handbook of Biological Physics, Vol. 4, edited by F. Moss and S. Gielen (North-Holland, 2001) pp. 619 – 684.

[13] M. Opper and O. Winther, in *Advanced mean field methods: theory and practice*, edited by M. Opper and D. Saad (MIT Press, Cambrdige, MA, 2001) pp. 7–21.

[14] J. Negele and H. Orland, *Quantum many-particle systems* (Perseus books, 1998).

[15] A. Kholodenko, J. Stat. Phys. **58**, 357 (1990).

[16] S. Osindero and G. Hinton, in *Advances in Neural Infor. Proces. Syst.*, Vol. 20 (2008); G. Taylor and G. Hinton, in *Proc. of ICML* (ACM, 2009) pp. 1025–1032.

[17] T. Plefka, J. Phys. A: Math. Gen. **15**, 1971 (1981); G. Biroli, **32**, 8365 (1999); Y. Roudi and J. Hertz, J.

Stat Mech: Theory and Exp. (2011).
[18] H. J. Kappen and J. J. Spanjers, Phys. Rev. E **61**, 5658 (2000).
[19] J. Tyrcha and J. Hertz, to be published (2013).
[20] C. De Dominicis, Phys. Rev. B **18**, 4913 (1978); H. Sompolinsky and A. Zippelius, Phys. Rev. Lett. **47**, 359 (1981).
[21] K. Fischer and J. A. Hertz, *Spin Glasses* (Cambridge University Press, 1991).